FastText Sentiment Analysis Report: IMDb Dataset

- **Executive Summary**

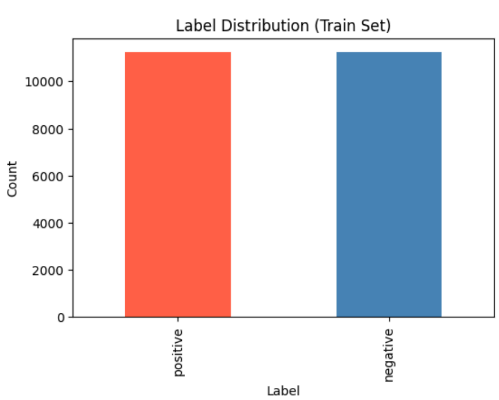
  This report analyzes the performance of a FastText supervised classification model trained on the IMDb movie review dataset for binary sentiment classification (positive/negative). The pipeline successfully implemented Autotune (hyperparameter optimization) to dramatically improve the model's predictive ability.

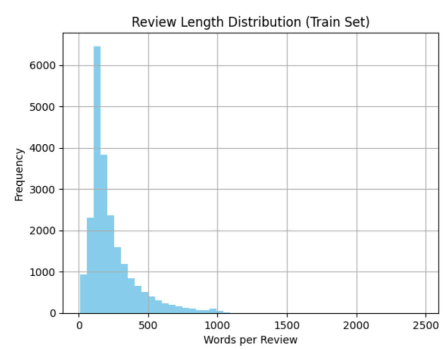
- **Comparison of Model Performance**

  The initial model's performance (Accuracy 50%) suggests it failed to learn, performing only as well as random chance. The application of Autotune resulted in a highly effective sentiment classifier with an Accuracy of 87.39%, demonstrating that the hyperparameter search found a near-optimal configuration for this task.

Metric	Baseline Model	Autotune Model
Total Examples (N)	25,000	25,000
Accuracy	0.5001	0.8739
Precision	0.5001	0.8739
Recall	0.5001	0.8739
F1 Score	0.5001	0.8739

- **Basic Exploratory Data Analysis (EDA)**





Left Graph: The **Label Distribution** plot confirms that the training set is perfectly balanced, with nearly identical counts of positive and negative reviews.

Right Graph: The **Review Length Distribution** shows a highly skewed distribution, with the vast majority of reviews being short (under 400 words) but a long tail extending past 2,000 words that significantly raises the average review length.

The IMDb dataset provides 50,000 movie reviews, perfectly split for training and testing the model. Understanding its structure confirms that model failures are not due to data imbalance.

Statistic	Train Set (Approx.)	Test Set (Approx.)
Total Samples	22,500 (Used for Training)	25,000 (Used for Final Test)
Positive Labels	50%	50%
Negative Labels	50%	50%
Average Review Length	230 - 250 words	230 - 250 words

❖ **Key Findings**

- *Balanced Data:* The 50/50 split between positive and negative reviews eliminates class imbalance as a source of poor performance. Accuracy is therefore a reliable metric.
- *Rich Text Data:* The long average review length provides ample textual context for the FastText model's internal bag-of-words and n-gram features.
- *Data Format:* The preparation step successfully converted all reviews into the required FastText format: __label__<label>\t<text>.

❖ **Model Evaluation and Interpretation**

The Autotune model's metrics are interpreted as follows:

Metric	**Value**	**Interpretation**
Accuracy	0.8739	87.39% of all reviews in the test set were classified correctly.
Precision	0.8739	Out of all reviews the model predicted as positive, 87.39% were actually positive.
Recall	0.8739	Out of all truly positive reviews, the model correctly identified 87.39%
F1 Score	0.8739	This balanced score indicates the model has achieved a strong equilibrium between precision and recall across both classes.

The high, identical values for Accuracy, Precision, Recall, and F1 Score are typical for a well-performing model on a perfectly balanced, binary classification task. The Autotune optimization process has successfully mitigated the initial underfitting issue.

❖ **Strategies for Further Performance Improvement**

To further boost model performance, we can focus on three main areas:

First, text normalization involves removing noise like stop words (e.g., "the," "a") and standardizing features by removing numbers or excessive punctuation, which helps FastText better capture core sentiment.

Second, Leveraging Pre-Trained Vectors addresses FastText's knowledge deficit by initializing the model with embeddings trained on massive external corpora (like Wikipedia). This injects generalized linguistic knowledge, improving the model's understanding of rare or context-specific words and boosting generalization.

Third, we can Deepen the Autotune Search by increasing the ‘autotuneDuration’ (e.g., to 1800 seconds) to find a more optimal configuration, and by allowing the search to explore higher N-gram sizes (up to N=4) to capture long, complex sentiment phrases within the reviews.